Supervised Learning Approaches to Classify Sudden Stratospheric Warming Events

CHRISTIAN BLUME* AND KATJA MATTHES*

Helmholtz Centre Potsdam, German Research Centre for Geosciences (GFZ), Potsdam, and Institute for Meteorology, Free University of Berlin, Berlin, Germany

ILLIA HORENKO

Institute of Computational Science, University of Lugano (USI), Lugano, Switzerland

(Manuscript received 8 July 2011, in final form 16 January 2012)

ABSTRACT

Sudden stratospheric warmings are prominent examples of dynamical wave–mean flow interactions in the Arctic stratosphere during Northern Hemisphere winter. They are characterized by a strong temperature increase on time scales of a few days and a strongly disturbed stratospheric vortex. This work investigates a wide class of supervised learning methods with respect to their ability to classify stratospheric warmings, using temperature anomalies from the Arctic stratosphere and atmospheric forcings such as ENSO, the quasi-biennial oscillation (QBO), and the solar cycle. It is demonstrated that one representative of the supervised learning methods family, namely nonlinear neural networks, is able to reliably classify stratospheric warmings. Within this framework, one can estimate temporal onset, duration, and intensity of stratospheric warming events independently of a particular pressure level. In contrast to classification methods based on the zonal-mean zonal wind, the approach herein distinguishes major, minor, and final warmings. Instead of a binary measure, it provides continuous conditional probabilities for each warming event representing the amount of deviation from an undisturbed polar vortex. Additionally, the statistical importance of the atmospheric factors is estimated. It is shown how marginalized probability distributions can give insights into the interrelationships between external factors. This approach is applied to 40-yr and interim ECMWF (ERA-40/ERA-Interim) and NCEP–NCAR reanalysis data for the period from 1958 through 2010.

1. Introduction

The variability of the north polar stratospheric vortex is a key dynamical feature of the middle atmosphere (Labitzke and van Loon 1999; Andrews et al. 1987)—specifically, its breakdown during winter resulting in a sudden stratospheric warming (SSW) (Scherhag 1952) taking place every 2 yr on average (Labitzke and Naujokat 2000). Obtaining insight into the dynamics, frequencies, and climatologies of stratospheric warming events is crucial to understand the underlying physical processes (Matsuno 1971; McIntyre 1982; Baldwin and Holton 1988) and relationships to atmospheric variability factors.

There have been several methods proposed in the past that can classify stratospheric warmings and measure the variability of the stratospheric vortex. Very common is the method based on the zonal-mean zonal wind at 60°N and 10 hPa originally introduced by the Stratospheric Group Berlin (Labitzke and Naujokat 2000) and incorporated by the World Meteorological Organization (WMO). It was used by Charlton and Polvani (2007) to compile climatologies of sudden stratospheric warmings derived from reanalyses data. It is a simple and effective method for measuring if and when a sudden stratospheric warming takes place leading to a vortex breakdown. Another method is based on the northern annular mode (NAM) (Baldwin and Dunkerton 2001) computed from geopotential anomalies in the middle stratosphere. The NAM measures the deviation from the climatological mean state of the polar middle atmosphere. It

* Current affiliation: Helmholtz Centre for Ocean Research Kiel (GEOMAR), Kiel, Germany.

Corresponding author address: Christian Blume, Helmholtz Centre Potsdam, German Research Centre for Geosciences (GFZ), Section 1.3: Earth System Modelling, Telegrafenberg, A20 329, 14473 Potsdam, Germany.
E-mail: christian.blume@met.fu-berlin.de

DOI: 10.1175/JAS-D-11-0194.1

measures the amount of disturbance but cannot alone be used to detect the occurrence of a vortex breakdown. It is widely used to detect downward propagation into the troposphere. The method based on 2D moments (e.g., Mitchell et al. 2011a) is a different way of measuring vortex variability. In contrast to the zonal wind measure and the NAM, it directly examines the geometrical structure of the vortex, such as position and size. In addition, it is used to measure the vortex strength.

In this work, a method is proposed that extends and combines the zonal wind measure and the NAM approach but does not examine the vortex structure. It incorporates significant atmospheric forcings, called external factors, that play an important role in the wintertime evolution of the polar stratosphere. These externals factors are the quasi-biennial oscillation (QBO; e.g., Holton and Tan 1980, 1982), the El Niño–Southern Oscillation (ENSO; e.g., Manzini et al. 2006), and the 11-yr solar cycle (SC; e.g., Gray et al. 2010). These forcings interact and create a complex and nonlinear dynamical response (e.g., Calvo et al. 2009; Richter et al. 2011). There are previous efforts, such as those of Labitzke and Kunze (2009a), Camp and Tung (2007a,b), and Mitchell et al. (2011b), that statistically investigated the impact of these forcings on the evolution of the polar vortex. Their analysis methods are linear, incorporating only a few factors at the same time. In this work, we use a nonlinear method with three external factors simultaneously to classify not only sudden stratospheric warmings but also minor and major final warmings as well as undisturbed vortex states at the same time. The classification procedure is a continuous analysis of stratospheric warming events for 52 consecutive winters in the period from 1958 through 2010.

In contrast to previous methods, the proposed classification method does not lead to a yes/no criterion but a continuous probability measure, which has the advantage of detecting the amount of deviation from the climatological mean state of the Arctic stratosphere. This disturbance of the polar vortex can then end up in one of the aforementioned stratospheric warming events. Dealing in terms of probabilities has the advantage of obtaining a temporal evolution of the likelihood of occurrence of a stratospheric warming state (e.g., major warming state), given the remaining states. In addition, the temporal onset, duration, and intensity of stratospheric warming events are calculated independently of a particular pressure level.

In this work, a wide class of supervised learning methods is considered and a classification method for stratospheric warmings based on a nonlinear statistical model, a neural network, is proposed. A supervised statistical method needs fixed pairs of input and output objects presented to it during training, meaning that the true outcome is known a priori. We show that a nonlinear model is suited better to recognize the complex nonlinear response between atmospheric forcings and polar vortex variability. Moreover, it is demonstrated that the approach based on a neural network can classify not only major midwinter stratospheric warmings (referred to hereafter as *major warmings*), but also minor stratospheric warmings (referred to hereafter as *minor warmings*), as well as major final stratospheric warmings (referred to hereafter as *final warmings*). So-called Canadian warmings will be grouped into the class of minor warming events.

Major and final warmings are characterized by a strong anomalous temperature increase at most pressure levels of the Arctic middle stratosphere, accompanied by a breakdown of the polar vortex and a reversal of the zonal stratospheric flow in midlatitudes from westerlies to easterlies. Major warmings are often preceded by blocking situations in the troposphere over the Atlantic and/or Pacific sector (Martius et al. 2009). Major warmings happen on average every other year during midwinter; hence there is enough time for the polar vortex to recover after a major warming. The polar vortex does not recover after a final warming as they take place at the transition between winter and summer circulation. Please note that final warmings naturally happen every year whereas final warmings (Labitzke and Naujokat 2000) in this work have to be accompanied by an anomalous temperature increase with respect to a climatology (*major final warming*). Minor warmings are characterized by an anomalous temperature increase and do not lead to a reversal of the zonal stratospheric flow in midlatitudes but rather to a disturbed polar vortex. Minor warmings often take place more than once in a given winter and are typically more upper-stratospheric events. Canadian warmings are minor warmings with the difference that anomalous temperatures are observed mainly in lower levels of the polar stratosphere. They are characterized by an enhancement of the Aleutian high (Labitzke and van Loon 1999).

This work is arranged as follows: Section 2 gives an overview of the data and input factors and introduces the calculation of the training sample. Section 3 reviews the supervised learning approaches and compares them with respect to their ability of classifying stratospheric warmings. Section 4 introduces the multilayer perceptron and estimates an optimal model architecture. Section 5 presents resulting warming probabilities and corresponding postprocessing in the 40-yr and interim European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40/ERA-Interim) and National Centers for Environmental Prediction

(NCEP)–National Center for Atmospheric Research (NCAR) reanalyses. The classification performance is evaluated and the impact of external factors computed. Section 6 classifies stratospheric warmings and presents a pathway toward understanding nonlinearities between different atmospheric forcings. Finally, conclusions and suggestions for further research are given.

2. Data

Three reanalysis datasets are utilized in this study: ERA-40 (Uppala et al. 2005), available up to 1 hPa from 1957 to 2002; ERA-Interim (Simmons et al. 2006), available up to 0.1 hPa from 1989 to the present; and the NCEP–NCAR reanalysis (Kalnay et al. 1996), available up to 10 hPa from 1948 to the present. The ERA-40 and ERA-Interim datasets resolve the entire stratosphere whereas the NCEP–NCAR reanalysis resolves only the middle and lower stratosphere. We will hereafter refer to the NCEP–NCAR reanalysis dataset as NCEP.

Time series from ERA-40 and ERA-Interim have been combined into one dataset that we refer to hereafter as ERA. This combination is justified by a small approximately Gaussian residual with zero mean calculated from the overlapping period (1989–2002) between the time series used in this work, separately calculated from ERA-40 and ERA-Interim. In this combined set, ERA-40 data have been used until 1 March 1989 and ERA-Interim data thereafter. This date has been selected because stratospheric temperatures and winds are very similar at and around this date, leading to a smooth transition between the two datasets. Both datasets, ERA and NCEP, are utilized for the time from 1 October 1958 through 1 May 2010, which covers 52 winters. ERA is utilized to train the statistical model. NCEP is utilized for validation because it is quite different from ERA in the polar region on account of its sparseness of observations, especially on the daily scale and for the presatellite era (Labitzke and Kunze 2005; Charlton and Polvani 2007). Also, it only reaches up to 10 hPa, leading to potentially different variability compared to ERA. ERA and NCEP have many input factors in common but, especially during the presatellite era, forcings in sea surface temperature (e.g., as seen in ENSO) along with equatorial stratospheric winds (e.g., as seen in the QBO) are significantly different.

Except for the zonal wind, all time series are normalized to ensure similar magnitudes according to

$$\hat{x}_t = (x_t - \mu_x)/\sigma_x \quad \forall t, \tag{1}$$

where x_t denotes the time series at time index t, μ_x is the mean of x, and σ_x is its standard deviation. In the literature this may also be called standardization. By applying Eq. (1), the normalized time series have zero mean and a variance of one.

a. The external factors: QBO, ENSO, and the solar cycle

This analysis makes use of three external factors that describe large-scale phenomena important for the stratosphere. It has been shown in previous studies (e.g., Labitzke and van Loon 1988; Camp and Tung 2007a,b; Mitchell et al. 2011b) that there exists a complex link among the external factors, namely the QBO, ENSO, SC, and the vortex variability. These studies showed, for example, that the least-perturbed vortex state is solar minimum and QBO west. It was also shown that the polar vortex is more likely to break down during El Niño–like conditions. Other studies have shown that this link is nonlinear (Calvo et al. 2009; Richter et al. 2011). The idea behind this work is to incorporate these external factors to classify stratospheric warmings on the one hand, and on the other to obtain insight into their statistical importance and interrelationships. In the following, we give a brief description of the corresponding indices.

The QBO index is the 50-hPa zonal-mean zonal wind anomaly averaged between 5°S and 5°N. For a representation of ENSO, we use the Niño-3.4 index (Trenberth 1997), which is the area-weighted average in sea surface temperature anomalies in the box from 170° to 120°E and from 5°S to 5°N. As a proxy for the solar irradiance, the radio flux at a wavelength of 10.7 cm (F10.7; ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SOLAR_RADIO/FLUX/Penticton_Observed/daily/DAILYPLT.OBS) is utilized. There are a few missing values in F10.7 that were filled by a linear interpolation, computed from the neighboring measurements. To reduce short-term fluctuations, the daily external factors have been sent through a low-pass filter calculating the 10-day running mean. This value was chosen to be more than a few days but much less than a month. Hence, daily short-term extremes are avoided in ENSO, QBO, and SC, but an approximately weekly resolution is retained.

b. Temperature

The classification procedure uses stratospheric temperatures because warming events have to be detectable naturally in Arctic temperatures. Temperature time series are considered at 10-, 20-, and 30-hPa levels where stratospheric warmings are always observed. They are also observable in upper and lower parts of the polar stratosphere depending on whether the event is a vortex split or displacement event. Vortex-splitting events tend to be observable near-instantaneously throughout most

parts of the polar stratosphere (~20–40 km) whereas vortex displacement off the pole increases with altitude (Matthewman et al. 2009).

The temperature time series are taken as an area-weighted average over the north polar cap between 60° and 90°N and are anomalies relative to their individual long-term mean. This treatment makes the time series equivalent to the northern annual mode in temperature (not geopotential) at the respective levels. The resulting temperature time series are highly correlated; however, the inputs to a classification approach should be decorrelated. A principal component analysis (PCA; von Storch and Zwiers 2001) of the three time series reveals that the first principal component (PC1) explains more than 90% of the overall variance in both ERA and NCEP (not shown). Therefore, PC1 is solely used as a robust representation of the temperature anomalies in the Arctic middle stratosphere, not favoring a particular pressure level. The normalized PC1 is displayed in Fig. 1 for ERA and NCEP, for a sample period of five winters from autumn 2005 to spring 2009. Because of the high degree of explained variance, PC1 is not used for classification only but also as a measure for the intensity of a stratospheric warming event (see section 6). Intensity is therefore a measure not only of strength but also of vertical expansion. It is taken as the maximum PC1 value during a warming event.

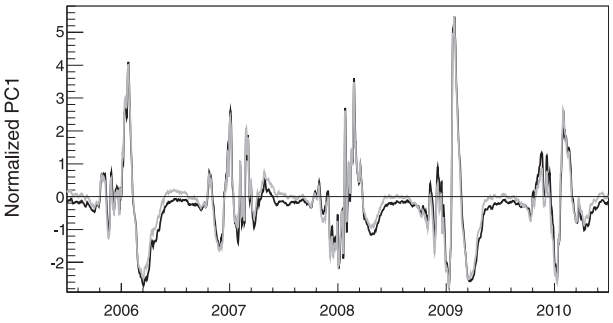

FIG. 1. Normalized first principal component for ERA-40 (black) and NCEP (gray) for a sample period from summer 2005 to summer 2010 covering five winters. Labeled is 1 Jan of the particular year.

c. Training sample

The main property of a supervised statistical model is that fixed sets of input and output objects are presented to it during training. The output is often called the *truth*, which has to be obtained externally. In our case, we make use of, among others, the zonal-mean zonal wind at 60°N and 10 hPa to receive time series of the four vortex states. We call the *training sample* the set of data that is presented to the statistical model during training. The statistical model will learn from the training sample. It can then be evaluated without using anything but the polar-cap temperature and the external factors. Here, the model learns the patterns between the given inputs and the vortex variability, making it possible to classify warming events in frequency, intensity, and duration and also to learn about impacts and relationships of the input factors.

Four classification time series have been produced, representing four different states of the Arctic stratosphere. The first three are major W^{major}, minor W^{minor}, and final W^{final} stratospheric warmings. The last is the undisturbed state W^{undis} in which no stratospheric warmings take place. Please note that W^{undis} does not denote that the polar vortex is not perturbed at all. It simply denotes the absent of stratospheric warming events. Three time series are used to calculate the training sample: PC1, the zonal-mean zonal wind at 60°N and 10 hPa $U^{10,60\text{N}}$; and the long-term mean of the 30-hPa polar-cap temperature $\overline{T^{30}}$.

The training sample is computed from ERA as follows. First, we define the disturbed state W^{dis} at time t as

$$W_t^{\text{dis}} := \begin{cases} 1 & \text{PC1}_t > 1 \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

which means that PC1 needs to exceed one standard deviation. This only happens during wintertime. The value of one sigma is relatively robust toward the number of derived major warmings and leads to just the right amount of average minor and final warmings per winter compared to observations. The undisturbed state W^{undis} is now given by

$$W_t^{\text{undis}} := 1 - W_t^{\text{dis}}, \tag{3}$$

which denotes the state that is least disturbed. Please note that the polar stratosphere is constantly perturbed by the dissipation of planetary waves (Labitzke and van Loon 1999). The next task is to extract major, minor, and final warmings from W^{dis}. We start with final warmings. The term $\overline{T_t^{30}}$ denotes a temporal measure so that $\overline{T_t^{30}} < 0$ represents the winter and $\overline{T_t^{30}} > 0$ the summer period ($\overline{T_t^{30}}$ is normalized). Therefore, values close to zero represent the transition between winter and summer or vice versa.

To classify major final warmings (referred to simply as *final warmings*), we have found the following definition to be appropriate:

$$W_t^{\text{final}} := \begin{cases} 1 & W_t^{\text{dis}} = 1 \wedge \overline{T_t^{30}} \geq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

which implies that disturbed states that happen in the transition phase from winter to summer are counted as final warming events. There are no disturbed states at the transition from summer to winter.

To determine major warmings, $U^{10,60N}$ needs to be incorporated. According to, for example, Charlton and Polvani (2007), a major warming event takes place if $U^{10,60N} < 0$ (easterlies) during the wintertime. Therefore, we define the *major warming state* as

$$W_t^{\text{major}} := \begin{cases} 1 & W_t^{\text{dis}} = 1 \wedge W_t^{\text{final}} = 0 \wedge U_t^{10,60N} < 0 \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

We added a neighborhood of 5 days in which the zonal-mean zonal wind can become easterly. The peak in temperature in the middle stratosphere during a major warming is usually a few days earlier than the wind reversal in 60°N. The *minor warming state* is now simply given by

$$W_t^{\text{minor}} := \begin{cases} 1 & W_t^{\text{dis}} = 1 \wedge W_t^{\text{final}} = 0 \wedge W_t^{\text{major}} = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

During the procedure of computing an appropriate training sample, it was ensured that contiguous events in W^{dis} were assigned to only one type of warming state. The warming states fulfill the condition

$$W_t^{\text{major}} + W_t^{\text{minor}} + W_t^{\text{final}} + W_t^{\text{undis}} = 1 \quad \forall t. \tag{7}$$

For instance, the results for the winter 1987/88 are displayed in Fig. 2. The minor, major, and final warming events are observed clearly. The time axis labels indicate the first day of the particular month in a given year.

d. Memory

Since there might be certain memory in the system, we need to get an estimate of the temporal lags of the external factors (QBO, ENSO, SC) that minimize the classification error. For reasons of simplicity and to reduce computational efforts, we restricted this calculation to a linear classification procedure (see next section) and only one target. This target has been chosen to be W^{major} as major warmings are of greatest interest.

A temporal lag larger than zero for SC does not seem to reduce the classification error at all. Therefore, the SC lag has been fixed to zero, and only the lags for QBO and ENSO have been varied between 0 and 180 days. An analysis with a step size of 1 day has been performed to find the optimal lags of 93 days for the QBO and 140 days for ENSO. These lags minimize the classification error and are used in all further analysis steps.

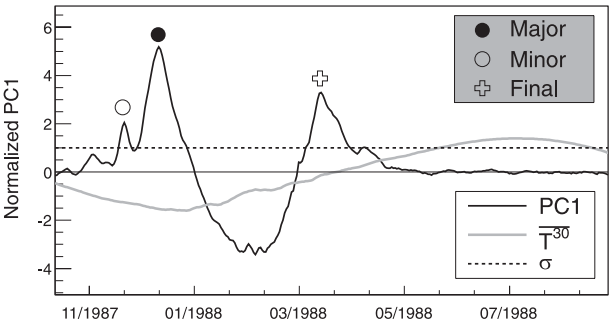

FIG. 2. Normalized PC1 for the winter 1987/88. The long-term mean of the 30-hPa temperature $\overline{T^{30}}$, the standard deviation $\sigma = 1$, and the estimated stratospheric warming events are displayed. Labeled is the first day of the particular month in a given year (MM/YYYY).

After estimating a set of lags for the external factors, it is interesting to calculate linear correlations between all input time series. It is generally favorable to use uncorrelated input variables when facing classification problems. The correlation matrix (not shown) reveals that there is no correlation apparent between any of the input variables. This also holds when keeping all time series at zero lag.

3. Statistical methods for classification

This section shortly reviews the three statistical methods that are later compared with respect to their classification performance to find the optimal method. Supervised learning (Theodoridis and Koutroumbas 2006; Marques de Sá 2001) is the task of deriving a function from a known training dataset consisting of pairs of input and output objects. For classification tasks the derived function is called a *classifier*.

Linear discriminant analysis (LDA) and linear support vector machines (LSVMs) represent the group of linear classifiers in our analysis, whereas multilayer perceptrons (MLPs) are generally nonlinear classifiers. In the following, it is assumed that there is a feature vector $\mathbf{x} \in \mathbb{R}^m$ and n training events. For reasons of simplicity, only two target classes (0, 1) are considered in the comparison of the three methods.

a. Linear discriminant analysis

Linear discriminant analysis (Wilks 1995; Montgomery et al. 2006) classifies data using a linear model. The discriminant function

$$y(x) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \beta_0 \tag{8}$$

is linear in its parameters $\boldsymbol{\beta} \in \mathbb{R}^m$, where $\beta_0 \in \mathbb{R}$ denotes a bias term that is usually selected so that $y < 0$ for class 0 and $y \geq 0$ for class 1. The equation for estimating $\boldsymbol{\beta} \in \mathbb{R}^{m+1}$ is

$$Y = \mathbf{X}\boldsymbol{\beta}, \tag{9}$$

where $Y \in \{0, 1\}^n$ and $\mathbf{X} \in \mathbb{R}^{n \times (m+1)}$. Applying the method of least squares, the normal equations of the classification problem are given by

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}Y \Leftrightarrow \boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}Y, \tag{10}$$

where $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ denotes the *Moore–Penrose pseudo inverse* of $\mathbf{X}$, which requires $\mathbf{X}$ to have full rank. LDA also assumes that the resulting residual is Gaussian distributed.

Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two events on the decision boundary. It follows that $y(\mathbf{x}_1) = y(\mathbf{x}_2) = 0$ and hence $(\mathbf{x}_1 - \mathbf{x}_2)^{\mathrm{T}}\boldsymbol{\beta} = 0$. Geometrically speaking, LDA is the task of finding a vector $\boldsymbol{\beta}$ that is orthogonal to the decision hyperplane.

b. Linear support vector machines

Support vector machines (Vapnik 1995; Burges 1998) try to find an optimal hyperplane that classifies data points by separating these points as much as possible. Let us assume that the data are not perfectly separable, which means that there will be a certain amount of misclassification. Then, a vector $\mathbf{w} \in \mathbb{R}^m$, a parameter $b \in \mathbb{R}$, and $\xi_i \geq 0$ can be found so that

$$y_i(\mathbf{x}_i^{\mathrm{T}}\mathbf{w} + b) - 1 + \xi_i \geq 0 \quad \forall i \in \{1, \ldots, n\}, \tag{11}$$

where the pair $(\mathbf{w}, b)$ defines the separating hyperplane. In Eq. (11), ξ_i denotes the so-called slack variable that measures the amount of misclassification of the feature vector $\mathbf{x}_i$. The classification margin $m = 2/|\mathbf{w}|$ is to be maximized with respect to the constraints given in Eq. (11). Hence, maximizing the margin is equivalent to minimizing the *cost function*

$$\mathbf{W} = \frac{1}{2}|\mathbf{w}|^2 + C\sum_{i=1}^{n}\xi_i, \tag{12}$$

with $\mathbf{w}$ subject to Eq. (11). The training events $\mathbf{x}_i$ that lie on the margin are called the support vectors (SVs). Also, $C \in \mathbb{R}$ in Eq. (12) denotes a parameter describing the trade-off between maximizing the margin and misclassification. Introducing slack variables is equivalent to support vector machines with soft margins.

Equation (12) is a constrained quadratic optimization problem that has a unique solution. It is solved by translating into Lagrangian formalism. The resulting nonzero Lagrangian multipliers define the support vectors.

In practice, there are only rare cases in which $\xi_i = 0\ \forall i$. Therefore, we are usually confronted with selecting parameter C. This is usually done empirically by trial and error, choosing the value of C that leads to the best generalization performance.

c. Multilayer perceptrons

Multilayer perceptrons (Bishop 1995; Ripley 1996) are fully connected feed-forward neural networks with one or more hidden layers located between input and output layer. Each layer consists of a certain number of neurons in parallel. Each neuron calculates a weighted linear combination of its N inputs so that its output y is given by

$$y = f\left(\sum_{i=1}^{N} w_i x_i + \theta\right), \tag{13}$$

where $w_i \in \mathbb{R}$ and $\theta \in \mathbb{R}$ denote weights and biases, respectively. Therefore, the weights are given at each synapse (connection between two neurons) and the biases at each neuron. The scalar function f in Eq. (13) is called the *transfer function* and is mostly (and also here) chosen to be a sigmoid of the form $f(x) = (1 + e^{-x})^{-1}$. The transfer functions at the output layer are chosen to be linear in our analysis.

Classification and generalization performance of an MLP stem from the nonlinear transfer functions and the numerous connections within the hidden layer(s). An MLP with a single hidden layer implements a single hyperplane. An MLP with two hidden layers implements arbitrary convex regions containing intersections of hyperplanes. It has been shown that an MLP with sigmoidal transfer functions and two hidden layers can approximate any continuous function (Kurkova 1992). For this reason, we will restrict our analysis to an MLP with a maximum of two hidden layers.

The learning algorithm used to determine the free parameters of the network is the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Avriel 2003), which is a faster variation of the standard back propagation. In our analysis, 1000 training iterations (epochs) are performed where it is made sure that the BFGS algorithm converges.

BFGS uses a gradient search technique to iteratively adjust weights and biases via minimizing a cost function given by

$$E = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{q}(y_{ij} - y_{ij}^{*})^2, \tag{14}$$

TABLE 1. Performance measures for LDA, LSVM, and MLP. The largest value (best performance) for the particular class and the particular performance measure in boldface.

	S			ε_S			I_{ROC}		
Class	LDA	LSVM	MLP	LDA	LSVM	MLP	LDA	LSVM	MLP
Major	0.857	0.814	**0.864**	0.490	0.542	**0.945**	0.984	0.981	**0.987**
Minor	0.831	0.733	**0.851**	0.080	0.107	**0.935**	0.963	0.941	**0.983**
Final	0.822	0.527	**0.862**	0.044	0.095	**0.950**	0.953	0.801	**0.997**
Undisturbed	0.898	**0.909**	0.882	0.869	**0.985**	0.981	0.995	0.998	**0.999**

where q denotes the number of classes. In Eq. (14), $y \in \{0, 1\}^q$ denotes the desired output and $y^* \in \mathbb{R}^q$ the actual MLP response.

The minimization problem is unconstrained and generally nonconvex. The effect of resulting local minima can be reduced by performing several training realizations with different initial values for weights and biases. It should be noted that the performance of an MLP may decrease significantly if the number of input dimensions becomes too large whereas methods such as SVM or LDA do not suffer from this problem to that extend. Here, however, the number of input dimensions is five and is therefore very small with respect to the number of training events.

d. Comparing the classification methods

The goal here is to compare the previously introduced statistical methods with respect to their classification performance for stratospheric warmings. For reasons of simplicity, only two classes are considered at a time. To compare the classification results, we briefly review three performance measures. Let y_S and y_B be the area-normalized response distributions for signal (class 1) and background (class 0), respectively.

1) The separation S between signal and background is given by

$$S = \frac{1}{2}\int_{-\infty}^{\infty} \frac{[y_S(x) - y_B(x)]^2}{y_S(x) + y_B(x)}\, dx. \tag{15}$$

2) The signal efficiency ε_S at a given background efficiency ε_B is defined by

$$\varepsilon_S = \int_a^{\infty} y_S(x)\, dx, \tag{16}$$

where a is given by $\varepsilon_B = \int_a^{\infty} y_B(x)\, dx$. A representative background efficiency of 0.01 has been selected.

3) The integral of the receiver operating characteristic (ROC) curve is given by

$$I_{\mathrm{ROC}} = \int_0^1 (1 - \varepsilon_B)\, d\varepsilon_S, \tag{17}$$

where $1 - \varepsilon_B$ is called the *background rejection*.

The three performance measures S, ε_S, and I_{ROC} are bounded between 0 and 1, where 0 means the worst and 1 the best performance achievable. Overviews of signal analysis can be found in Fawcett (2006) and Spackman (1989).

The tuning parameters for LSVM and MLP have been chosen somewhat intuitively for this comparison (LDA does not have tuning parameters). For LSVM, the cost parameter C was varied between 0.1 and 10 and the value with the best performance ($C = 1$) was selected for further analysis. For the MLP, we chose 10 neurons in the first and 5 neurons in the second hidden layer. These values are of the same order as the number of inputs to avoid overfitting. The MLP was trained 10 times with different, randomly chosen initial parameters and the realization with the best performance was kept. The training for each method was performed in such a way that events where assigned alternating to train and test datasets.

The classification results are presented with respect to the aforementioned performance measures in Table 1 for LDA, LSVM, and MLP. The largest value (best performance) is underlined for the particular class and performance measure. First, the MLP clearly outperforms the linear models in all performance measures when classifying stratospheric warmings. Out of the linear models, LSVM performs better than LDA for the warming classes with respect to ε_S but worse with respect to S and I_{ROC}. If the goal is to only discriminate between undisturbed and disturbed states, LSVM is even slightly better than MLP. This is not unexpected since the only difference between a disturbed and undisturbed Arctic stratosphere is a simple linear cut on PC1 [see Eq. (2)]. In this work, we are particularly interested in the correct classification of stratospheric warmings. Hence, MLP clearly wins this method comparison with respect to the given performance measures. MLP seems to be able to classify stratospheric warmings rather well as all performance measures are close to one. Hence, MLP is our method of choice for the following

analysis. In the next section, the MLP analysis is explained in greater detail and a pathway toward an optimal MLP architecture is presented.

4. Multilayer perceptrons and model architecture

Neural networks are widely used methods for efficient pattern recognition (Ripley 1996). Here, an artificial neural network recognizes patterns in temperature anomalies and external factors to classify stratospheric warming events as major, minor, and final warmings. More specifically, the neural network here is an MLP in which all neurons of a certain layer are connected via synapses to all neurons in the neighboring layers (see section 3). An MLP is one of the most general and best understood neural network types (Bishop 1995). As in section 3, we use the BFGS learning algorithm to determine weights and biases. The training is performed in such a way that events are assigned in alternation to train and test datasets.

The input layer consists of five input neurons, which are $\overline{T^{30}}$, PC1, QBO, ENSO, and SC. The output layer consists of four neurons representing four different states of the polar stratosphere. The first three are major, minor, and final stratospheric warmings. The last is the undisturbed state, in which no stratospheric warmings take place.

The optimal model architecture of the MLP is estimated. The number of hidden layers as well as the number of hidden neurons within these layers needs to be determined. The dimensions of input and output layers have been specified in section 2. Each MLP setting is considered to be a separate statistical model.

We are making use of methods from information theory that were shown to have remarkable ability to discriminate between statistical models (Burnham and Anderson 2002). In comparison with cross-validation (Kohavi 1995), this approach is computationally much less expensive and leads to the model setting with the best descriptive power whereas cross-validation focuses more on forecasting. As mentioned above, events are assigned alternating to train and test dataset, thereby incorporating a simple cross-validation with neighboring events that helps to avoid overfitting.

We start by reviewing an important information criterion. The Bayesian information criterion (BIC; Schwarz 1978; Priestley 1981) is given by

$$\mathrm{BIC} = N_T \cdot \ln(\sigma_e^2) + N_P \cdot \ln(N_T), \tag{18}$$

where N_T denotes the overall sample size, N_P the number of free parameters in the model, and σ_e^2 the variance of the residual distribution. This version of the BIC given in Eq. (18) is applicable under the assumption that the errors are independent and identically distributed according to a Gaussian distribution (Priestley 1981). This assumption holds for our problem (not shown). The number of free parameters of the MLP is given by

$$N_p = \sum_{i=1}^{M-1} m_i(m_{i+1} + 1) + m_M, \tag{19}$$

where m_i denotes the number of neurons in layer i and M the total number of layers in the MLP.

The BIC can be understood as an estimator for the balance between explained variance and the number of free model parameters. The model with the smallest information criterion of all tested models is the preferred model. Hence, the BIC differences can be defined as

$$\Delta_i = \mathrm{BIC}_i - \mathrm{BIC}_{\min}, \tag{20}$$

where $\mathrm{BIC}_{\min}$ denotes the minimal BIC value within the sample of tested models and i one model out of this sample ($\Delta = 0$ for the best model).

To determine the optimal model architecture, the MLP needs to be trained many times with different model configurations. The MLP training has been repeated 10 times with different, randomly chosen initial parameters for each model configuration. To reduce the effect of local minima, the resulting σ_e^2 used to calculate the BIC is taken as the mean of those 10 optimizations.

The number of hidden neurons is varied in the hidden layers. The results of Eq. (20) are displayed in Fig. 3 where the white square indicates $\Delta = 0$. This procedure was repeated using the Akaike information criterion (Akaike 1974), which led to a more complicated model architecture with significantly more free parameters and was therefore rejected. The resulting optimal model setting has two hidden layers with 23 neurons in the first and 4 neurons in the second layer. The MLP has now been trained 100 times with this specific architecture. The run with the smallest error is chosen. The classification results of this run are presented in the following sections.

5. Probabilities of stratospheric warmings

In this section first classification results based on conditional probabilities for each of the classes are presented. Additionally, the statistical method is validated. The following results are based on the multilayer perceptron as described in the previous section.

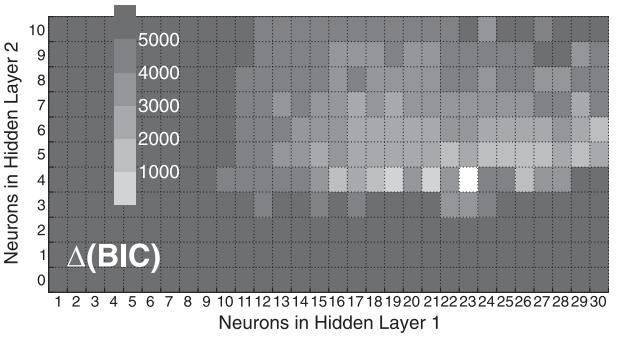

FIG. 3. The BIC differences according to Eq. (20) for varying number of hidden neurons. The white square ($\Delta = 0$) denotes the optimal model architecture with 23 neurons in the first and 4 neurons in the second hidden layer. Note the valley of small BIC values around the optimum.

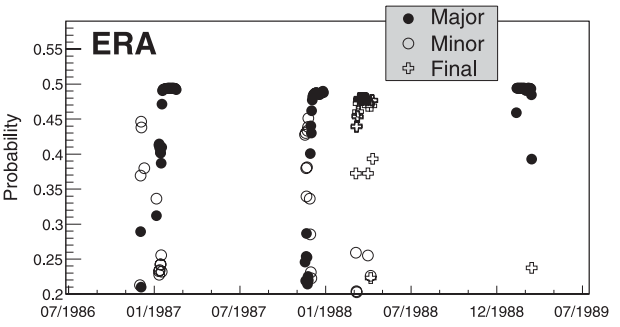

FIG. 4. Evolution of the probabilities in ERA for major, minor, and final warmings for the three winters in the period from summer 1986 to summer 1989. Labeled is the first day of the particular month in a given year (MM/YYYY).

To ensure that the MLP response values can be interpreted as conditional probabilities, the value y_i of output neuron i needs to be transferred via

$$p_i = \frac{\exp(y_i)}{\sum_{j=1}^{n} \exp(y_j)}, \tag{21}$$

which is known as the *softmax function* (Ripley 1996), and has the useful property

$$p_i \in [0,1] \quad \forall i \quad \text{and} \quad \sum_{i=1}^{n} p_i = 1. \tag{22}$$

Having computed conditional probabilities, we are interested in determining a threshold value for each class above which a certain probability is significantly different from the background. We will call this the cut_i at class i. To do so we integrate over the area-normalized background probability distribution $P_{B,i}$ for each class i such that

$$\alpha = \int_0^{\text{cut}_i} P_{B,i}(x)\, dx, \tag{23}$$

where $\alpha = 0.999$, so that a probability p_i greater than cut_i is significantly different from the background at a confidence level of 99.9%. We obtain 0.32, 0.34, and 0.25 for the major, minor, and final warming class, respectively. The cuts are relatively small, which indicates a good classification performance.

a. Three sample winters

We now want to obtain further insights into the MLP response. Three adjacent sample winters are selected that include all three types of warming events. Figure 4 shows the evolution of the probabilities for major, minor, and final warmings for the period from summer 1986 to summer 1989. Results for ERA (training dataset) and NCEP (validation dataset; not shown) are very similar. The winter 1987/88 appears to be the most variable on this period. A minor warming in November 1987, lasting about 5 days, is observed. A major warming (Baldwin and Dunkerton 1989; Naujokat et al. 1988) takes place in the beginning of December 1987, lasting about 20 days. A short minor warming appears as a precursor to this major warming. Ultimately, a final warming lasting about 15 days takes place in March (Labitzke and Naujokat 2000). The probabilities shown in Fig. 4 give a good representation of what was observed (cf. Fig. 2). The classification performance is now assessed in greater detail.

b. Classification performance

In addition to the performance measures introduced in section 3d, we compute the classification performance with respect to the mean difference MD_i for class i given by

$$\text{MD}_i = \frac{1}{N_T} \sum_{j=1}^{N_T} |p_{ij} - p_{ij}^*|, \tag{24}$$

where p_{ij} denotes the training value at output neuron i and sample index j and p_{ij}^* the corresponding MLP response. Also, N_T denotes the overall sample size representing the total number of steps in time. For a perfect classification it is expected that $\text{MD}_i = 0$ for all i.

Table 2 presents the performance measures for each class as calculated from the ERA classification results. A very high classification performance is obtained. The separation, the signal efficiency, and the integral of the ROC curve are very close to one for all classes. This represents a very good ability of discriminating signal from background events.

TABLE 2. The performance measures for the optimal MLP setting for each class.

Class	S	ε_S ($\varepsilon_B = 0.01$)	I_{ROC}	MD
Major	0.992	0.969	0.996	0.0023
Minor	0.985	0.929	0.990	0.0026
Final	0.962	0.968	0.987	0.0015
Undisturbed	0.990	0.991	0.995	0.0010

TABLE 3. Relative impact (%) on the MLP response according to Eq. (25) for each input neuron and output class.

Input	Major	Minor	Final	Undisturbed
$\overline{T^{30}}$	22.2	22.4	54.0	0.8
PC1	26.9	23.7	31.2	97.5
QBO	19.3	19.9	4.6	0.7
ENSO	17.4	17.9	4.9	0.6
SC	14.3	15.9	5.3	0.4

MD is very close to zero for all classes, which implies that only in rare cases the MLP response is not close to the data that the MLP has been trained with (the training sample; see section 2). Therefore, the MLP is able to reliably detect major, minor, and final warming states and, of course, the undisturbed state.

c. Impact of the input neurons

It is of great interest to estimate the individual impact of the five input neurons on the MLP response. This gives an insight into the statistical importance of each of the input factors. The impact $I_{i,k}$ of input factor i on output class k is simply chosen to be the variance of MLP response differences given by

$$I_{i,k} = \mathrm{Var}(y_k^{(i)} - y_k), \tag{25}$$

where y_k denotes the MLP response at output neuron k and $y_k^{(i)}$ the corresponding MLP response where the input factor i was set to zero. If an input neuron had no impact on the MLP, Eq. (25) would give zero. Table 3 presents the relative impact in percent on the MLP response according to Eq. (25) for each input neuron and output class.

It is observed that the impacts are quite different for different output classes. For the undisturbed case only PC1 plays an important role. This is expected as the undisturbed state is simply defined by a linear cut on PC1 (see section 2c). The final warming state is mostly governed by PC1 and $\overline{T^{30}}$ since the definition of the final warming state was only based on these two factors (see section 2c).

When looking at major and minor warming states, the external factors become more important and necessary to discriminate major from minor warmings. The QBO shows the largest impact, followed by ENSO and the solar cycle, in agreement with previous studies (e.g., Labitzke and Kunze 2009b; Camp and Tung 2007a,b; Mitchell et al. 2011b) that also investigated the impact of these forcings and found a similar ranking. Hence, the neural network combines QBO, ENSO, and SC in a nonlinear fashion to distinguish between major and minor stratospheric warmings. Therefore, the external factors, namely QBO, ENSO, and SC, should be incorporated in order to classify stratospheric warmings successfully. It was mentioned earlier that there is practically no linear correlation between any of the input time series. However, as Table 3 shows, there exist nonlinear combinations of input factors that lead to different stratospheric warming states.

6. Stratospheric warming climatologies

This section presents stratospheric warming climatologies extracted from resulting probabilities for 52 winters from 1958 through 2010. To identify stratospheric warmings, we need to define a threshold above which a signal in one of the output neurons is counted as an event signal. An event signal has to be significant; hence, it needs to exceed the cut values (see section 5). To get an estimate for the training dataset ERA, we calculated the first derivative dQ/dp of the cumulative density function of the response distribution of each warming class. As an increasing derivative denotes a regime change, we define the thresholds where dQ/dp starts rising from its constant level with increasing quantiles. The resulting thresholds for ERA are 0.41, 0.41, and 0.45 for major, minor, and final warming events, respectively. We have found that the resulting ERA warming event numbers and distributions are not sensitive with respect to slightly different thresholds.

As the validation set NCEP is a priori unknown, and to avoid counting events caused by a possible overfitting, we need to find a reasonable NCEP threshold that is larger than any of the ERA thresholds but smaller than the theoretical limit given by Eq. (22). An NCEP threshold of 0.47 for all warming classes was selected leading to reasonable distributions and event numbers as presented in the following. The resulting NCEP events are more sensitive with respect to this threshold than the ERA events but can still be changed in the percentage range and the event numbers and monthly distributions would not change significantly.

a. Warming events

To obtain stratospheric warming events, we need to group contiguous warming days. To do so, minimal

TABLE 4. Total number of stratospheric warming events and relative number of events per year for the different warming classes and the two datasets. The uncertainties are given in parentheses (standard error of mean).

Data	Major	Minor	Final	Total
Total				
ERA	31	74	27	132
NCEP	26	76	28	130
Relative				
ERA	0.6 (0.1)	1.4 (0.2)	0.5 (0.1)	2.5 (0.3)
NCEP	0.5 (0.1)	1.5 (0.2)	0.5 (0.1)	2.5 (0.3)

temporal distances between adjacent warming events need to be defined. If this distance is exceeded without output neuron i above the given probability threshold, then warming event i is finished and a new warming event may eventually take place. For these distances we choose 30 days for major warmings, 5 days for minor warmings, and 100 days for final warmings. The number for final warmings is rather arbitrary as they may only take place once a year during the transition from winter to summer circulation. We selected 30 days for major warmings because it is known from observations (Labitzke and van Loon 1999) that major warmings may last 20 days but that neighboring major warmings in the same winter are at least one month apart. The relatively short period of 5 days for minor warmings was chosen since minor warmings are usually not preceded by a great cooling in the Arctic stratosphere, as major warmings are (Labitzke and van Loon 1999; Charney and Drazin 1961). Therefore, adjacent minor warming events can be closer than major warming events.

First results of this procedure are shown in Table 4 for ERA and the validation dataset NCEP. The absolute number (upper part) and relative number (lower part) of warming events are presented. It is observed that values for ERA and NCEP are very similar for all warming classes. This indicates a successful validation of the classification procedure using NCEP. Only the major warming case shows slightly fewer events in NCEP than in ERA. This discrepancy for major warmings has also been reported by Charlton and Polvani (2007).

To summarize, there is a major warming event approximately every other year whereas minor warmings happen at least once a year on average. Major final warmings take place every second year, too. These results are in good agreement with Charlton and Polvani (2007), who find approximately 0.6 SSWs per winter. Labitzke and Naujokat (2000) find approximately 0.5 major midwinter warmings, approximately 1 minor warming (half of which are Canadian warmings), and 0.25 major final warmings per winter. The differences between ERA and NCEP found in our work are due to differences in the two datasets, particularly in PC1 and ENSO during the presatellite era before 1979.

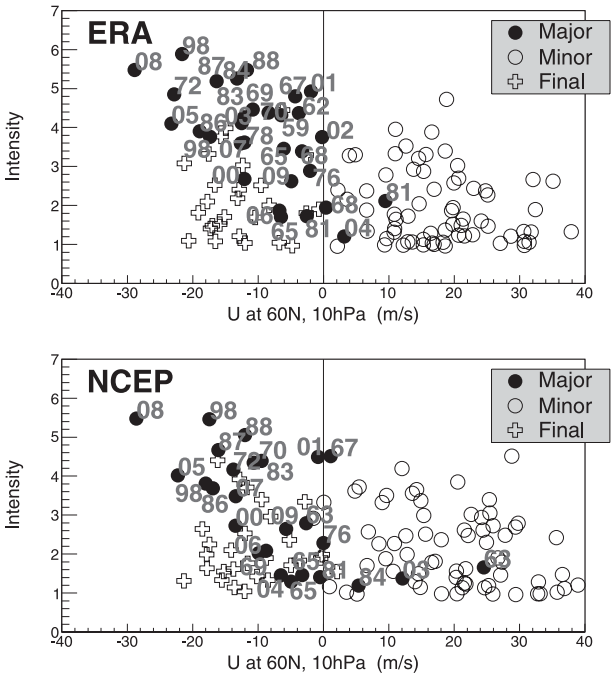

FIG. 5. Scatter diagram of stratospheric warming intensity against the zonal-mean zonal wind at 60°N and 10 hPa. Numbers represent winters (e.g., 98 denotes the winter 1998/99) in which major warming(s) took place. The results are shown for (top) ERA and (bottom) NCEP. Values for minor warmings temporally very close to major or final warmings are not shown as they lead to ambiguous wind results.

b. Change in circulation

The question remains whether the detected major and final warming events lead to a vortex breakdown and therefore a change in circulation (easterly zonal winds) in the stratosphere in midlatitudes. Minor warmings should only slow down the circulation but not reverse it. To tackle this question, the zonal-mean zonal wind at 60°N and 10 hPa is incorporated. If the zonal wind is negative (easterlies), then a change in circulation took place and the polar vortex broke down. An interval of 20 days around the central warming date of major and final warmings was considered to find the minimum zonal wind.

The result of this analysis is shown in Fig. 5 for all warming classes and both datasets. Values for minor warmings temporally very close to major or final warmings are not shown as they lead to ambiguous wind results. The numbers represent the winter in which a major warming took place (e.g., 98 denotes the winter 1998/99). The zonal wind reversed for almost all major and final warming events in ERA and NCEP, which confirms the classification procedure.

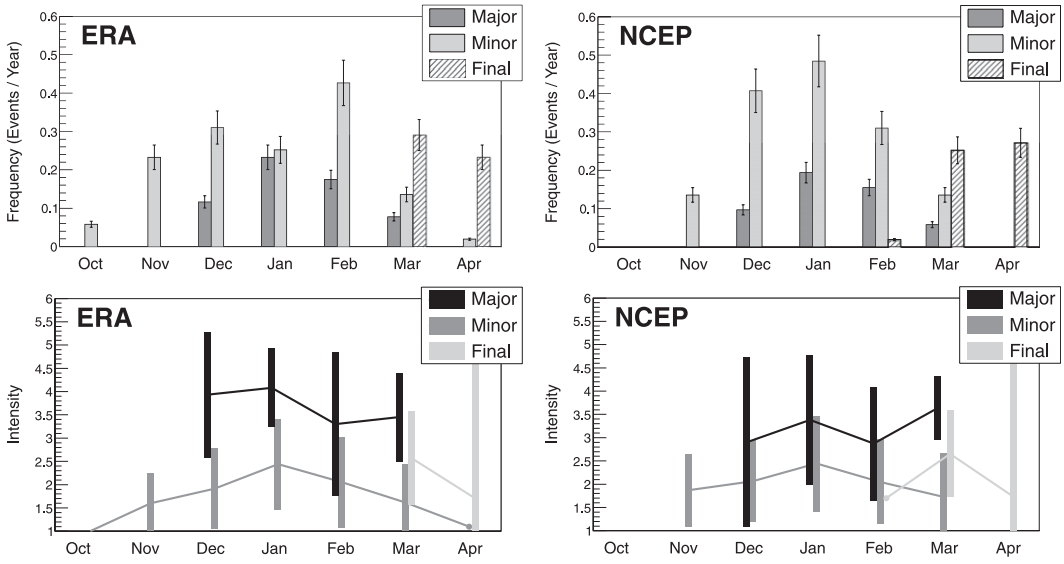

FIG. 6. Monthly distributions in (top) events per year and (bottom) intensity of the three warming classes for (left) ERA and (right) NCEP. The error bars represent (top) the standard error of mean and (bottom) the standard deviation. Please note the different shading schemes for frequency (histograms) and intensity (graphs).

There are only a few clear falsely detected major warming events for which the vortex was disturbed and the circulation slowed down but did not reverse. In ERA these are the winters 1981/82 and 2004/05. In NCEP these are the winters 1963/64, 1984/85, and 2003/04. It is the nature of a statistical method that it is never 100% effective. However, all final warming events were classified correctly. Despite a few differences, the classified stratospheric warmings are in good agreement with previous studies (Charlton and Polvani 2007; Labitzke and Naujokat 2000). None of the detected minor warming events led to a change in circulation.

c. *Stratospheric warming frequencies*

The classification results are now analyzed and presented in more detail with respect to their occurrences and intensities. Monthly climatologies of major, minor, and final warmings are shown in Fig. 6. The uncertainties are displayed as error bars. First, the distributions for ERA and NCEP are similar. Most major warmings take place in January. Minor warmings happen all throughout the winter but most take place in February for ERA and January for NCEP, whereas final warmings clearly peak in March and April. There are no major warmings taking place in November, which is in agreement with observations (Labitzke and Naujokat 2000).

Major warmings show highest intensities with large variability followed by minor and final warmings. As expected, the minor warming intensities peak in January and decrease toward beginning and end of the winter. The final warming intensities are also very variable and peak in March.

Charlton and Polvani (2007) show monthly distributions for major warmings retrieved from a classification method based on the zonal-mean zonal wind at 60°N and 10 hPa. These results are similar to the distribution for major warmings shown in Fig. 6. There have been a few SSWs found by Charlton and Polvani (2007) in November that were most likely Canadian warmings. They found more SSWs in March simply because some of those are counted as final warmings in our analysis.

It is of great interest to investigate the temporal evolution of the three warming classes over the 52-yr period. Their frequency of occurrence and intensity in bins of 4 yr is presented in Fig. 7. The frequency distributions resemble observations rather well (Labitzke and Naujokat 2000). For instance, the clear minimum of major warming activity observed in the 1990s is obtained. There are also periods of higher major warming activity in the 1970s. Minor warmings were especially frequent during the 1980s and 1990s. Final warmings do not show significant occurrence variabilities. The results for ERA and NCEP in Fig. 7 are again qualitatively similar. Differences appear mostly during the presatellite era before 1979. In comparison to Charlton and Polvani (2007), differences for the major warming case are mainly due to different methodologies and classification strategies.

The intensities presented in Fig. 7 are also similar in ERA and NCEP. We see a great decrease in major

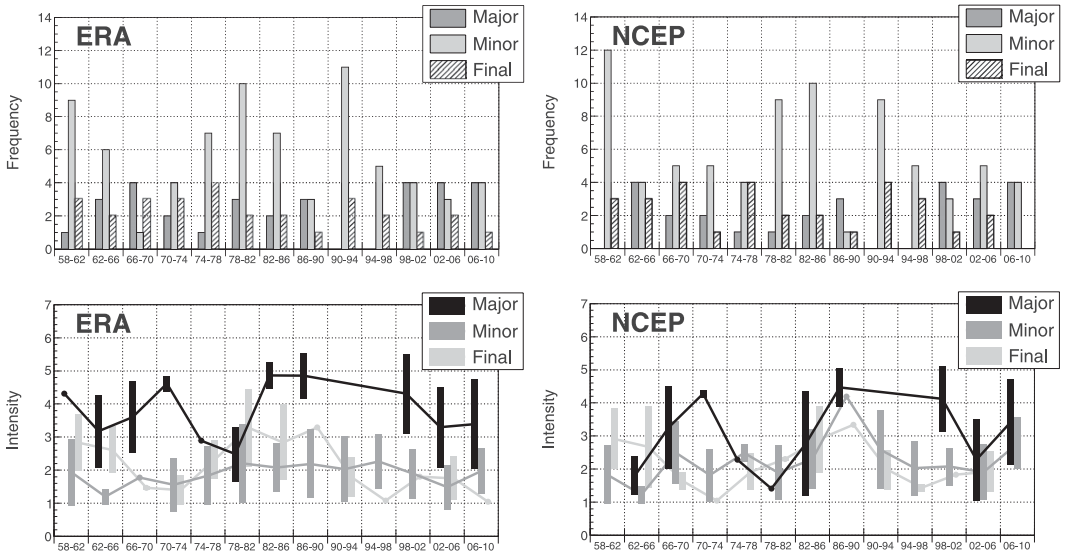

FIG. 7. Distributions in bins of 4 yr of (top) occurrence and (bottom) intensity of the three warming classes for (left) ERA and (right) NCEP. All bins start at 1 Sep and stop at 31 Aug of the respective years. Error bars represent one standard deviation around the mean. Note the different shading schemes for frequency (histograms) and intensity (graphs).

warming intensity during the 1970s and large intensities in the past 30 yr. The mean minor warming intensities seem to be rather constant throughout the whole period whereas the final warming intensities show a peak in the 1980s and then decrease to minor warming levels.

Mean intensity and the corresponding standard deviation of the three warming events and their duration in days for ERA and NCEP are presented in Table 5. The results for ERA and NCEP agree rather well. On average, major and final warmings last about 20 days and minor warmings only 8 days. There is a large variability in duration as the standard deviation takes values of about 10 days for each warming class. On average, major warmings are twice as intense as minor and final warmings with medium variability.

Table 5 also shows the linear correlation between intensity and duration for each warming class. All correlation factors are significant (t test) at the 95% confidence level. For ERA, all correlation factors are greater than 0.6, which leads us to the expected conclusion that warmings with larger intensities generally last longer, and vice versa. For NCEP, the correlation factors are slightly smaller.

d. Marginalized probability distributions

The neural network can be considered as a function (classifier) mapping from a five-dimensional input space to a four-dimensional probability space. To retain an understanding of the relationships between the input factors despite the high dimensionality, we are marginalizing the resulting probability distributions. Motivated by previous studies, we are particularly interested in the relationships among QBO, ENSO, and SC. Therefore, these factors have been varied and the resulting MLP response investigated.

PC1 has been fixed and the responses have been averaged for the midwinter between December and February. Additionally, the results have been split for solar maximum and solar minimum conditions where a value of 120 solar flux units (sfu) of the F10.7 solar radio flux was used to separate the two regimes. The resulting marginalized probability distributions are shown in Fig. 8 for the major warming state. The shading denotes the probability of the occurrence of a major warming and the black thick line an approximately significant probability of 0.3. The numbers in Fig. 8 represent the winter

TABLE 5. Mean intensity (standard deviations) of stratospheric warming events and their mean duration (days) for the different stratospheric warming events in ERA and NCEP. The corresponding standard deviation is given in parentheses. The correlation between duration and intensity is also given. All correlation factors are significant (t test) at the 95% confidence level.

Data	Class	Intensity	Duration	Correlation
ERA	Major	3.7 (1.2)	23.0 (10.7)	0.61
	Minor	1.9 (0.9)	8.4 (8.3)	0.75
	Final	2.2 (1.0)	20.1 (10.9)	0.67
NCEP	Major	3.2 (1.4)	16.7 (12.3)	0.53
	Minor	2.1 (1.0)	8.9 (10.5)	0.52
	Final	2.2 (0.9)	20.2 (10.5)	0.41

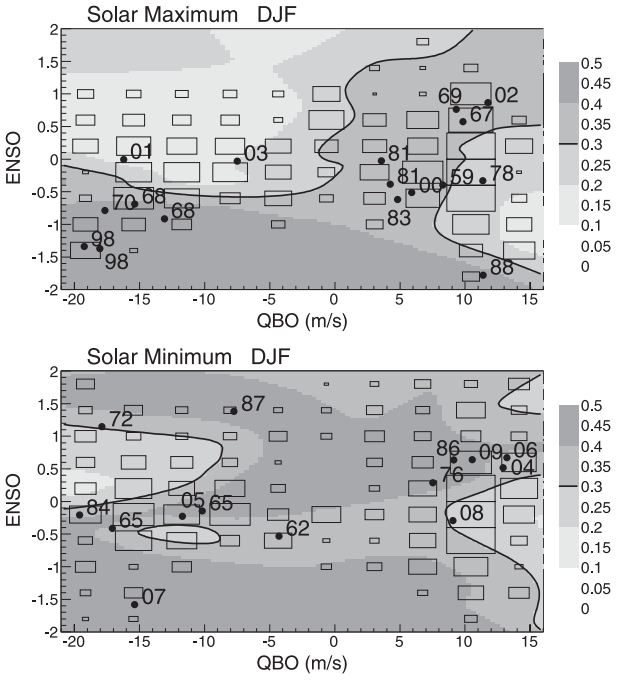

FIG. 8. Marginalized probability distributions (shading; black line denotes $p = 0.3$) for the major warming state depending on ENSO and the QBO for (top) solar maximum and (bottom) solar minimum for PCI = 3, denoting a moderate vortex disturbance. The darker the shading is, the higher the probability for a major warming. The numbers denote the winter of a major warming (e.g., 87 denotes the winter 1987/88). The superimposed boxes represent a schematic frequency distribution of QBO and ENSO; the larger the box is, the greater the population density for a particular bin (i.e., a large box stands for a high frequency of a particular combination of QBO–ENSO values).

in which the major warming took place (e.g., 87 denotes the winter 1987/88).

A moderately high disturbance of PCI = 3 (cf. Fig. 5) was selected, implying that the condition for the disturbed state is fulfilled and the MLP discriminates between two states: major warmings and minor warmings. Because of the aforementioned averaging, the patterns in Fig. 8 represent climatological mean states for DJF. These patterns are highly nonlinear, which emphasizes the usefulness of a nonlinear statistical method. Previous studies have shown that the considered forcings interact and create a complex and nonlinear dynamical link (e.g., Calvo et al. 2009; Richter et al. 2011).

The superimposed boxes in Fig. 8 represent a schematic frequency distribution of QBO and ENSO: the larger the box, the greater the population density of a particular bin (i.e., a large box stands for a high frequency of a particular combination of QBO–ENSO values and vice versa). Hence, these population densities are naturally different for solar maxima and minima. Highly populated regions are observed, but also combinations of QBO and ENSO that have not been seen in the data at all. The larger the population is, the more we can trust the MLP response. In regions with zero population (no boxes), the MLP predicts probabilities. Considering the good validation results for NCEP, which is an unseen dataset, we believe that these MLP predictions are trustworthy. Nevertheless, they need to be confirmed by data from chemistry–climate model simulations.

There are two main regions that are not populated. The first is the region of large negative ENSO values (La Niña) and small absolute QBO values (around zero) for both solar maximum and minimum. La Niña events are rather rare and the transition between QBO west and QBO east and vice versa is very fast (often within a month), whereas a QBO phase (east or west) can last about a year. The other underpopulated region is that of large positive ENSO values (El Niño) during solar maximum for almost all values of the QBO. Hence, El Niño events are only rarely found during solar maximum conditions.

Figure 8 presents various probability features for major warmings. Despite the averaging, almost all major warming events fall into the significant area of $p \geq 0.3$, indicating a robust classification. Please note the aforementioned averaging over SC regimes and the midwinter, implying that probabilities for a specific event may be different from what is shown in Fig. 8. There are regions of high probabilities for QBO west and solar maximum conditions, as also found by Labitzke and Kunze (2009b) and Camp and Tung (2007a). However, there is a region for strong QBO west in both solar maximum and minimum, in which moderate and La Niña–like ENSO events show only small probabilities. The high population density in this region makes the probabilities particularly trustworthy. This indicates strong nonlinear relationships between QBO and ENSO as also found by, for example, Calvo et al. (2009) and Richter et al. (2011). Linear interrelationships, as emphasized by, for instance, Camp and Tung (2007a,b) and Labitzke and Kunze (2009b), are not sufficient to explain this pattern.

The very intense major warming of the winter 2008/09 (solar minimum, QBO west, and slightly negative ENSO values) is very close to the significant region in Fig. 8. Hence, this major warming along with the major warming in 2006/07 is part of the nonlinear rules determined by the MLP, whereas these events have been previously treated as exceptions from linear rules (e.g., Labitzke and Kunze 2009a).

Despite the high probabilities, only a few major warmings are found to happen during the transition from QBO west to QBO east or vice versa (see Fig. 8). This is because of the aforementioned fast transition between

QBO phases (west ↔ east). Moreover, it is known from observations (Baldwin et al. 2001) that the QBO phase transition takes place mostly during the Northern Hemisphere summer. By definition, sudden stratospheric warmings take place only during the wintertime.

During QBO east and solar maximum conditions, only negative ENSO values show significant probabilities. During QBO west, moderate and El Niño–like ENSO conditions lead to significant major warming probabilities. For solar minimum and QBO east, strong positive ENSO events lead to large probabilities, too. A probability minimum is observed for ENSO values close to zero. This minimum appears also for QBO west but for slightly negative ENSO values and is more dependent on the strength of the QBO. In general, the probability for a disturbance to become a major warming leading to a vortex breakdown is greater during solar minimum conditions (note the large significant area) than during solar maximum. As also found by Butler and Polvani (2011), El Niño–like and La Niña–like conditions make the occurrence of major stratospheric warmings more likely as opposed to neutral ENSO conditions. The only exception is the small major warming probability for El Niño–like conditions during solar maximum and QBO east.

7. Conclusions

This work classifies stratospheric warmings by considering Arctic stratospheric temperature anomalies together with atmospheric forcings (or external factors) that influence the polar vortex, namely the QBO, ENSO, and the solar cycle (SC). The classification procedure is applied to data from the ERA-40/ERA-Interim (jointly referred to as ERA) and the NCEP–NCAR (herein simply NCEP) reanalysis for 52 consecutive winters from 1958 to 2010. Optimal lags of the external factors are determined using linear discriminant analysis.

Three supervised learning approaches (LDA, LSVM, MLP) are introduced and compared with respect to their ability to classify stratospheric warmings. It is shown that the nonlinear MLP outperforms the linear methods (Table 1). This is in agreement with previous work showing that the external factors nonlinearly influence the polar vortex evolution (e.g., Calvo et al. 2009; Richter et al. 2011). The MLP is therefore used as the method of choice to classify stratospheric warmings in major, minor, and major final warming events. This approach extends and combines the zonal wind measure and the NAM approach applied in previous studies. It incorporates the polar-cap temperature and significant external factors simultaneously leading to a continuous probability measure, indicating the amount of deviation from the climatological mean state.

It is shown how an appropriate training sample (Fig. 2) can be calculated. Using this training sample, the optimal MLP architecture is determined using methods from information theory (Fig. 3). Using various performance measures, the classification procedure is successfully validated (Table 2). It is shown how resulting stratospheric warming probabilities (Fig. 4) are postprocessed.

The statistical impact of the input factors on the individual output classes is computed (Table 3). It is shown that the atmospheric variability factors are an essential part of the classification procedure as they discriminate between minor and major stratospheric warmings. They are less important for final warmings and show only a small impact on the undisturbed state (Table 3). Despite the absence of any linear correlations between the external factors, there are nonlinear combinations that help distinguish between warming classes. The QBO was found to have the largest impact, followed by ENSO and the solar cycle. This ranking was also found by previous work (e.g., Labitzke and Kunze 2009b; Camp and Tung 2007a,b; Mitchell et al. 2011b) that investigated the influence of these forcings on the polar vortex.

It is shown that detected major and final warming events lead to a vortex breakdown and a reversal of the zonal flow at 60°N (Fig. 5) except for a few cases (two in ERA, three in NCEP). Reasonable distributions of stratospheric warming events by month and year of occurrence and intensity are presented (Figs. 6 and 7), which are in agreement with previous work made by Charlton and Polvani (2007) and Labitzke and Naujokat (2000), who also compiled climatologies of stratospheric warming events. On average, major warmings show intensities that are twice as large as those of minor or final warmings. Final warmings last as long as major warmings but twice as long as minor warmings. We find largely positive significant correlations greater than 0.6 between intensity and duration of the warming events (Table 5).

Marginalized probability distributions depending on QBO and ENSO, for both solar maximum and solar minimum conditions, are presented (Fig. 8). The results contain the linear QBO–SC relationships presented by Camp and Tung (2007a) and Labitzke and Kunze (2009b). However, we show that the interrelationships between the external factors are nonlinear as previously suggested. QBO–SC relationships are nonlinearly modulated by ENSO (Calvo et al. 2009). It appears that El Niño–like conditions (Camp and Tung 2007b) during QBO west favor the occurrence of major warmings and vice versa during QBO east. This pattern is more prominent for solar maxima than for solar minima. For the solar minima, also El Niño–like conditions and QBO east point to large major warming probabilities. We find that

major warmings are generally more likely during solar minimum conditions. For the solar minima, there are only two regions that do not favor major warmings, which are small but positive ENSO values during QBO east and small but negative ENSO values during QBO west. This pattern also depends on the strength of the particular QBO phase. As also found by Butler and Polvani (2011), major warmings are more likely during El Niño–like and La Niña–like conditions as opposed to neutral ENSO conditions. An exception to this is only observed for El Niño–like conditions during solar maxima and QBO east. In addition, we show that the extraordinary major warming of the winter 2008/09 lies close to the significant climatological area that indicates a possible vortex breakdown. Therefore, this event is part of the nonlinear rules learned by the MLP. A three-dimensional animation through the winter of the probabilities indicated in Fig. 8 can be found online (http://nathan.gfz-potsdam.de/doc/sswanim.gif).

Several improvements of the current statistical framework are possible. The introduction of the geopotential height into the MLP input layer would further enhance the classification results, as it provides direct information about the polar vortex strength. Introducing a memory of 1 or 2 days would also improve the classification but exponentially increase computation time. Incorporating volcanic influences may also improve the classification procedure.

It is shown that a statistical model with the current set of input factors needs to recognize nonlinear patterns to reliably classify stratospheric warmings. However, there are not only neural networks that can cope with this challenge. One may also think of applying methods such as support vector machines with nonlinear kernels or nonlinear functional discriminant analysis.

The current framework will be applied to data from chemistry–climate model simulations to validate the current results and to investigate the difference of a data constrained model, such as reanalyses, to a free-running CCM. Since the relationships between the external factors and polar vortex variability are generally different in reanalyses and CCMs, the MLP has to be trained separately for each CCM. The generally nonlinear interrelationships along with various measures (frequency, intensity, duration, etc.) can then be compared among model simulations. Further application of our framework to the prediction of stratospheric warmings is also envisaged.

Acknowledgments. This work has been carried out within the Helmholtz University Young Investigators Group NATHAN funded by the Helmholtz Association through the President's Initiative and Networking Fund, the GFZ Potsdam, and by FU Berlin. We thank the members of the Stratosphere Group at the Institute for Meteorology of the FU Berlin for helpful discussions, particularly Karin Labitzke, Ulrike Langematz, Markus Kunze, and Anne Kubin. This work was partially supported by the Center for Scientific Simulations (FU Berlin, Project: ''Advanced Methods of Time Series Analysis and their Application to Climate Research and Insurance Risk Optimization'') and Swiss Platform for High-Performance and High Productivity Computing (HP2C). We are grateful to two anonymous reviewers for their constructive comments.

REFERENCES

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Autom. Control,* **19,** 716–723.

Andrews, D. G., J. R. Holton, and C. B. Leovy, 1987: *Middle Atmosphere Dynamics*. Academic Press, 489 pp.

Avriel, M., 2003: *Nonlinear Programming: Analysis and Methods*. Dover, 528 pp.

Baldwin, M. P., and T. J. Dunkerton, 1989: The stratospheric major warming of early December 1987. *J. Atmos. Sci.,* **46,** 2863–2884.

——, and J. R. Holton, 1988: Climatology of the stratospheric polar vortex and planetary wave breaking. *J. Atmos. Sci.,* **45,** 1123–1142.

——, and T. J. Dunkerton, 2001: Stratospheric harbingers of anomalous weather regimes. *Science,* **294,** 581–584.

——, and Coauthors, 2001: The quasi-biennial oscillation. *Rev. Geophys.,* **39,** 179–229.

Bishop, C., 1995: *Neural Networks for Pattern Recognition*. Oxford University Press, 482 pp.

Burges, C. J. C., 1998: A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discovery,* **2,** 121–167.

Burnham, K. P., and D. R. Anderson, 2002: *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 488 pp.

Butler, A. H., and L. M. Polvani, 2011: El Niño, la Niña, and stratospheric sudden warmings: A reevaluation in light of the observational record. *Geophys. Res. Lett.,* **38,** L13807, doi:10.1029/2011GL048084.

Calvo, N., M. A. Giorgetta, R. Garcia-Herrera, and E. Manzini, 2009: Nonlinearity of the combined warm ENSO and QBO effects on the Northern Hemisphere polar vortex in MAECHAM5 simulations. *J. Geophys. Res.,* **114,** D13109, doi:10.1029/2008JD011445.

Camp, C. D., and K.-K. Tung, 2007a: The influence of the solar cycle and QBO on the late-winter stratospheric polar vortex. *J. Atmos. Sci.,* **64,** 1267–1283.

——, and ——, 2007b: Stratospheric polar warming by ENSO in winter: A statistical study. *Geophys. Res. Lett.,* **34,** L04809, doi:10.1029/2006GL028521.

Charlton, A. J., and L. M. Polvani, 2007: A new look at stratospheric sudden warmings. Part I: Climatology and modeling benchmarks. *J. Climate,* **20,** 449–469; Corrigendum, **20,** 5551.

Charney, J. G., and P. G. Drazin, 1961: Propagation of planetary-scale disturbances from the lower into the upper atmosphere. *J. Geophys. Res.,* **66,** 83–109.

Fawcett, T., 2006: An introduction to ROC analysis. *Pattern Recog. Lett.,* **27,** 861–874.

Gray, L. J., and Coauthors, 2010: Solar influences on climate. *Rev. Geophys.,* **48,** RG4001, doi:10.1029/2009RG000282.

Holton, J. R., and H. C. Tan, 1980: The influence of the equatorial quasi-biennial oscillation on the global atmospheric circulation at 50 mb. *J. Atmos. Sci.,* **37,** 2200–2208.

——, and ——, 1982: The quasi-biennial oscillation in the Northern Hemisphere lower stratosphere. *J. Meteor. Soc. Japan,* **60,** 140–148.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–472.

Kohavi, R., 1995: A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc. Int. Joint Conf. on Artificial Intelligence, Vol. 2, Montreal, QC, Canada, IJCAI, 1137–1143. [Available online at http://www.cs.iastate.edu/~jtian/cs573/Papers/Kohavi-IJCAI-95.pdf.]

Kurkova, V., 1992: Kolmogorov's theorem and multilayer neural networks. *Neural Netw.,* **5,** 501–506.

Labitzke, K., and H. van Loon, 1988: Associations between the 11-year solar cycle, the QBO and the atmosphere. Part I: The troposphere and stratosphere in the Northern Hemisphere winter. *J. Atmos. Terr. Phys.,* **50,** 197–206.

——, and ——, 1999: *The Stratosphere: Phenomena, History, and Relevance*. Springer, 179 pp.

——, and B. Naujokat, 2000: The lower arctic stratosphere in winter since 1952. *SPARC Newsletter,* No. 15, SPARC Office, Toronto, ON, Canada, 11–14. [Available online at http://www.atmosp.physics.utoronto.ca/SPARC/News15/15_Labitzke.html.]

——, and M. Kunze, 2005: Stratospheric temperatures over the arctic: Comparison of three data sets. *Meteor. Z.,* **14,** 65–74.

——, and ——, 2009a: On the remarkable arctic winter in 2008/2009. *J. Geophys. Res.,* **114,** D00102, doi:10.1029/2009JD012273.

——, and ——, 2009b: Variability in the stratosphere: The Sun and the QBO. *Climate and Weather of the Sun-Earth System (CAWSES): Selected Papers from the 2007 Kyoto Symposium,* Terrapub, 257–278.

Manzini, E., M. Giorgetta, M. Esch, L. Kornblueh, and E. Roeckner, 2006: The influence of sea surface temperatures on the northern winter stratosphere: Ensemble simulations with the MAECHAM5 model. *J. Climate,* **19,** 3863–3881.

Marques de Sá, J. P., 2001: *Pattern Recognition: Concepts, Methods, and Applications*. Springer, 318 pp.

Martius, O., L. M. Polvani, and H. C. Davies, 2009: Blocking precursors to stratospheric sudden warming events. *Geophys. Res. Lett.,* **36,** L14806, doi:10.1029/2009GL038776.

Matsuno, T., 1971: A dynamical model of the stratospheric sudden warming. *J. Atmos. Sci.,* **28,** 1479–1494.

Matthewman, N. J., J. G. Esler, A. J. Charlton-Perez, and L. M. Polvani, 2009: A new look at stratospheric sudden warmings. Part III. Polar vortex evolution and vertical structure. *J. Climate,* **22,** 1566–1585.

McIntyre, M. E., 1982: How well do we understand the dynamics of stratospheric warmings? *J. Meteor. Soc. Japan,* **60,** 37–65.

Mitchell, D. M., A. J. Charlton-Perez, and L. J. Gray, 2011a: Characterizing the variability and extremes of the stratospheric polar vortices using 2D moment analysis. *J. Atmos. Sci.,* **68,** 1194–1213.

——, L. J. Gray, and A. J. Charlton-Perez, 2011b: The structure and evolution of the stratospheric vortex in response to natural forcings. *J. Geophys. Res.,* **116,** D15110, doi:10.1029/2011JD015788.

Montgomery, D., and Coauthors, 2006: *Introduction to Linear Regression Analysis*. 4th ed. Wiley-Interscience, 612 pp.

Naujokat, B., K. Labitzke, R. Lenschow, K. Petzoldt, and R.-C. Wohlfahrt, 1988: The stratospheric winter 1987/88: An unusually early major midwinter warming. *Beilage zur Berliner Wetterkarte,* SO 6/88, 20 pp.

Priestley, M. B., 1981: *Spectral Analysis and Time Series*. Academic Press, 890 pp.

Richter, J. H., K. Matthes, N. Calvo, and L. J. Gray, 2011: Influence of the quasi-biennial oscillation and El Niño–Southern Oscillation on the frequency of sudden stratospheric warmings. *J. Geophys. Res.,* **116,** D20111, doi:10.1029/2011JD015757.

Ripley, B. D., 1996: *Pattern Recognition and Neural Networks*. 1st ed. Cambridge University Press, 416 pp.

Scherhag, R., 1952: Die explosionsartigen Stratosphärenerwärmungen des Spätwinters 1951/52. *Ber. Deut. Wetterdienst,* **6,** 51–63.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.,* **6,** 461–464.

Simmons, A., S. M. Uppala, D. Dee, and S. Kobayashi, 2006: ERA-Interim: New ECMWF reanalysis products from 1989 onwards. *ECMWF Newsletter,* ECMWF, Reading, United Kingdom No. 110, 26–35.

Spackman, K. A., 1989: Signal detection theory: Valuable tools for evaluating inductive learning. *Proceedings of the Sixth International Workshop on Machine Learning,* Morgan Kaufmann, 160–163.

Theodoridis, S., and K. Koutroumbas, 2006: *Pattern Recognition*. 3rd ed. Elsevier, 837 pp.

Trenberth, K. E., 1997: The definition of El Niño. *Bull. Amer. Meteor. Soc.,* **78,** 2771–2777.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Vapnik, V., 1995: *The Nature of Statistical Learning Theory*. 1st ed. Springer-Verlag, 188 pp.

von Storch, H., and F. W. Zwiers, 2001: *Statistical Analysis in Climate Research*. 1st ed. Cambridge University Press, 484 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. 1st ed. International Geophysics Series, Vol. 59, Academic Press, 467 pp.